SA TR ISO/IEC 24028:2024
ISO/IEC TR 24028:2020

STANDARDS Australia

Technical Report

Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence

SA TR ISO/IEC 24028:2024

This Australian Technical Report was prepared by IT-043, Artificial Intelligence. It was approved on behalf of Standards Australia's Standards Development and Accreditation Committee on 11 April 2024.

This Technical Report was published on 03 May 2024.

The following are represented on Committee IT-043:

Australian Computer Society
Australian Healthcare and Hospitals Association
Australian Human Rights Commission
Australian Industry Group
Australian Information Industry Association
Australian Institute of Company Directors
Australian Institute of Health & Safety
Australian Securities and Investments Commission
CHOICE
Consult Australia
Consumers Federation of Australia
CSIRO
Ethics, AI and ADM Professional Group
Gradient Institute
Human Factors and Ergonomics Society of Australia
National Association of Testing Authorities Australia
NSW Data Analytics Centre
Queensland AI Hub
Royal Australian and New Zealand College of Radiologists
Therapeutic Goods Administration (TGA)
University of Melbourne
University of New South Wales
University of Technology Sydney
Western Sydney University

ISBN 978 1 76139 615 1

SA TR ISO/IEC 24028:2024
ISO/IEC TR 24028:2020

Technical Report

Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence

First published as SA TR ISO/IEC 24028:2024.

Preface

This Technical Report was prepared by the Standards Australia Committee IT-043, Artificial Intelligence.

The objective of this document is to survey topics related to trustworthiness in AI systems, including the following:

— approaches to establish trust in AI systems through transparency, explainability, controllability, etc.;

— engineering pitfalls and typical associated threats and risks to AI systems, along with possible mitigation techniques and methods; and

— approaches to assess and achieve availability, resiliency, reliability, accuracy, safety, security and privacy of AI systems.

The specification of levels of trustworthiness for AI systems is out of the scope of this document.

This document is identical with, and has been reproduced from, ISO/IEC TR 24028:2020, *Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence.*

As this document has been reproduced from an international document, a full point substitutes for a comma when referring to a decimal marker.

Australian or Australian/New Zealand Standards that are identical adoptions of international normative references may be used interchangeably. Refer to the online catalogue for information on specific Standards.

The terms “normative” and “informative” are used in Standards to define the application of the appendices or annexes to which they apply. A “normative” appendix or annex is an integral part of a Standard, whereas an “informative” appendix or annex is only for information and guidance.

Contents

Preface ii
Foreword v
Introduction vi
1 Scope 1
2 Normative references 1
3 Terms and definitions 1
4 Overview 7
5 Existing frameworks applicable to trustworthiness 7
5.1 Background 7
5.2 Recognition of layers of trust 8
5.3 Application of software and data quality standards 8
5.4 Application of risk management 10
5.5 Hardware-assisted approaches 10
6 Stakeholders 11
6.1 General concepts 11
6.2 Types 12
6.3 Assets 12
6.4 Values 13
7 Recognition of high-level concerns 13
7.1 Responsibility, accountability and governance 13
7.2 Safety 14
8 Vulnerabilities, threats and challenges 14
8.1 General 14
8.2 AI specific security threats 15
8.2.1 General 15
8.2.2 Data poisoning 15
8.2.3 Adversarial attacks 15
8.2.4 Model stealing 16
8.2.5 Hardware-focused threats to confidentiality and integrity 16
8.3 AI specific privacy threats 16
8.3.1 General 16
8.3.2 Data acquisition 16
8.3.3 Data pre-processing and modelling 17
8.3.4 Model query 17
8.4 Bias 17
8.5 Unpredictability 17
8.6 Opaqueness 18
8.7 Challenges related to the specification of AI systems 18
8.8 Challenges related to the implementation of AI systems 19
8.8.1 Data acquisition and preparation 19
8.8.2 Modelling 19
8.8.3 Model updates 21
8.8.4 Software defects 21
8.9 Challenges related to the use of AI systems 21
8.9.1 Human-computer interaction (HCI) factors 21
8.9.2 Misapplication of AI systems that demonstrate realistic human behaviour 22
8.10 System hardware faults 22
9 Mitigation measures 23
9.1 General 23
9.2 Transparency 23